Прикладная статистика
3. Критерии однородности, ANOVA.

Родионов Игорь Владимирович
rodionov@bigdatateam.org

Устойчивость к выбросам

Непараметрические, т.е. использующие только ранги наблюдений тесты всегда являются устойчивыми к выбросам. К таким тестам относятся критерий Манна-Уитни, критерий знаковых рангов Уилкоксона, критерий знаков.

Параметрические, т.е. основанные на самих значениях наблюдений тесты часто являются неустойчивыми к выбросам. Все тесты, использующие выборочные характеристики, являются неустойчивыми к выбросам. Примеры: t-критерий Стьюдента, критерий Фишера равенства дисперсий, критерий Аспина-Уэлча.

Определение нужного объема выборки

Как определить количество наблюдений, которое нужно, чтобы считать проверку некоторой гипотезы значимой? Для начала мы должны определить, какую ошибку можно считать незначимой.

Допустим, в t-критерии Стьюдента положим ошибку в 0.1 незначимой. Т.е. если $|\overline{X} - \mu| \leq 0.1$, то мы не будем отвергать гипотезу $H_0 : EX_1 = \mu$. Имеем, при $|\overline{X} - \mu| = 0.1$,

$$\sqrt{n}\frac{\overline{X} - \mu}{s} \approx t_{1-\alpha/2},$$

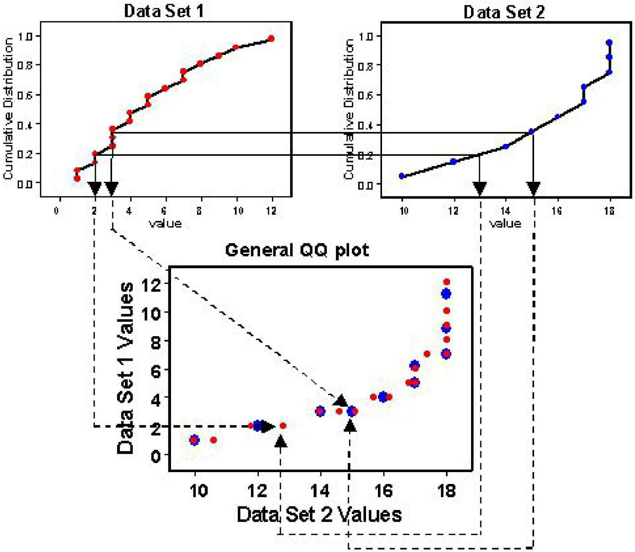
где $t_{1-\alpha/2}$ – квантиль $St(n-1)$. Зададим $\alpha = 0.05$, тогда при больших n $t_{0.975} \approx 2$. Будем считать, что $s \approx 1$, тогда $\sqrt{n} \approx \frac{2}{0.1} = 20$, откуда $n \approx 400$.

General QQ-plot

Пусть имеются две (независимые) выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ с функциями распределения F и G соответственно. Допустим, мы хотим проверить гипотезу $H_0 : F = G\left(\frac{x-a}{\sigma}\right)$.

General QQ-plot – это график, на который нанесены точки $\left(\widehat{F}_n^{-1}\left(\frac{j}{m}\right), Y_{(j)}\right)$ и $\left(X_{(i)}, \widehat{G}_m^{-1}\left(\frac{i}{n}\right)\right)$. Если точки лежат примерно на одной прямой, то гипотеза H_0 близка к верности.

General QQ-plot



Критерий Смирнова

Обсудим теперь критерии проверки двух выборок на однородность. Для решения этой задачи можно адаптировать критерии согласия, например, критерий Колмогорова-Смирнова.

Пусть $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ – две независимые выборки с непрерывными ф.р. F и G соответственно, а $\widehat{F}_n(x)$ и $\widehat{G}_m(x)$ – эмпирические функции распределения этих выборок. Определим

$$D_{n,m} = \sup_x |\widehat{F}_n(x) - \widehat{G}_m(x)|,$$

тогда при верной гипотезе $H_0 : F = G$ статистика $\sqrt{\frac{nm}{n+m}} D_{n,m}$ имеет табличное распределение. При $n, m \geq 20$ оно приближается распределением Колмогорова.

Общий критерий Андерсона-Дарлинга

Пусть $(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), \ldots, (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$ – k независимых выборок с функциями распределений $F_1, \ldots, F_k$ соответственно. Пусть $\widehat{F}_1, \ldots, \widehat{F}_k$ – эмпирические функции распределения этих выборок и $\widehat{H}_N(x), N = \sum_i n_i,$ – эмпирическая функция распределения общей совокупности наблюдений.

Тогда статистика

$$\Omega^2 = \sum_{i=1}^{k} n_i \int_{\mathbb{R}} \frac{(\widehat{F}_i(x) - \widehat{H}_N(x))^2}{\widehat{H}_N(x)(1 - \widehat{H}_N(x))} d\widehat{H}_N(x)$$

имеет табличное распределение при верной гипотезе $H_0 : F_1 = \ldots = F_k$.

Пусть имеются наблюдения признака X на $N = \sum_i n_i$ объектах.

Хотим проверить, зависят ли значения признака X (а точнее, его среднее) от некого фактора A, принимающего значения (уровни) $(A_1, \ldots, A_k)$.

Пусть при $A = A_j$ значения признака X заданы выборкой $\{X_{ij}\}_{i=1}^{n_j}$, $1 \leq j \leq k$.

Линейная (т.н. однофакторная) модель:

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

$i = 1, \ldots, n_j$, $j = 1, \ldots, k$.

μ – глобальное среднее признака X;
α_j – отклонение от μ, вызванное влиянием j-того уровня фактора A;
ε_{ij} – н.о.р. случайные ошибки, $E\varepsilon_{ij} = 0$.

Т.е. средние значения X во всех выборках одинаковы тогда и только тогда, когда $\alpha_1 = \ldots = \alpha_k$.

Критерий Фишера

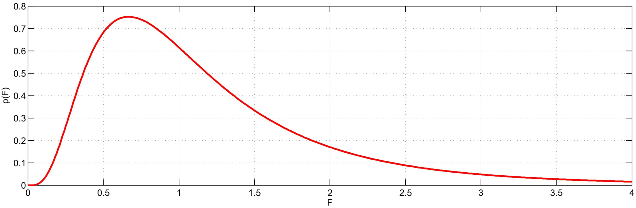
Для проверки гипотезы $H_0 : \alpha_1 = \ldots = \alpha_k$ против альтернативы $H_1 : H_0$ *неверна* используется статистика

$$F = \frac{\sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2} \cdot \frac{N-k}{k-1}.$$

В случае выполнения H_0 и предположений метода

$$F \sim F(k-1, N-k).$$

Критерий обычно выбирается правосторонним.

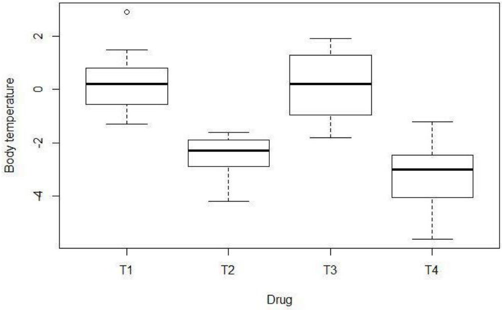


Предположения метода:

1. выборочные распределения средних значений признака во всех группах нормальны;
2. дисперсия значений признака во всех выборках одинакова;
3. наблюдения независимы.

- Первое предположение считается выполненным, если распределение признака во всех группах нормально, или если объёмы выборок примерно одинаковы и $N - k - 1 \geq 20$.
- Второе предположение считается выполненным, если отношение наибольшей выборочной дисперсии к наименьшей не превосходит 10.
- При $n_1 = \ldots = n_k$ метод устойчив к нарушению первых двух предположений.
- Если объёмы выборок различаются, нарушение предположения о равенстве дисперсий может привести к росту вероятности ошибки первого рода.
- Выбросы могут оказывать существенное влияние на результат.

Пример: исследуется эффективность четырёх жаропонижающих средств, в составе которых один и тот же активный ингредиент присутствует в разных дозировках. Для каждой из четырёх групп из 15 морских свинок известно изменение температуры после введения жаропонижающего. Есть ли различия в действии препаратов?



Критерий Фишера проверки гипотезы H_0 об отсутствии различий в действии препаратов дает p-value $= 5.43 \times 10^{-14}$.

Критерий Краскела-Уоллиса

Пусть $\{X_{ij}\}, 1 \le i \le n_j, 1 \le j \le k$ – независимые выборки с ф.р. $F_j(x) = F(x - \alpha_j)$. Проверим гипотезу об отсутствии сдвига $H_0 : \alpha_1 = \ldots = \alpha_k$ против альтернативы H_1 : H_0 *неверна*.

Пусть $R_{ij} = R(X_{ij})$ – ранг наблюдения X_{ij} в общей совокупности, $\overline{R}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ij}$, $\overline{R} = \frac{1}{N} \sum_{i,j} R_{ij} = \frac{N+1}{2}$.

Статистика критерия Краскела-Уоллиса

$$W = (N-1) \frac{\sum_{j=1}^{k} n_j (\overline{R}_j - \overline{R})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (R_{ij} - \overline{R})^2}$$

имеет табличное распределение (при верной H_0), которое при $n_j > 5 \ \forall j$ приближается распределением χ^2_{k-1}.

Критерий Джонкхиера

Данный критерий используется для проверки гипотезы $H_0 : \alpha_1 = \ldots = \alpha_k$ против альтернативы $H_1' : \alpha_1 \le \ldots \le \alpha_k$. Статистика критерия

$$S = \sum_{j=1}^{k} \sum_{i=1}^{n_j} a_{ij},$$

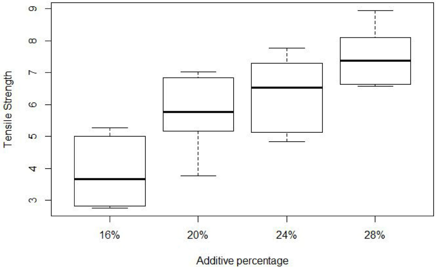
где a_{ij} – количество наблюдений из первых $j-1$ выборок, меньших X_{ij}. При верности гипотезы H_0 имеет табличное распределение.

На альтернативе H_1' критерий Джонкхиера имеет большую мощность, чем критерий Краскела-Уоллиса, чем и объясняется его использование.

Кроме того, оба этих критерия являются устойчивыми к наличию выбросов в данных.

Пример

Исследуется зависимость предела прочности армированного бетона с разной концентрацией укрепляющих добавок: 16, 20, 24 и 28%. Меняется ли средний предел прочности вместе с концентрацией добавок?



H_0: концентрация добавок не влияет на среднюю прочность.
H_1: концентрация добавок влияет на среднюю прочность (критерий Краскела-Уоллиса): p-value $= 0.0042$.
H_1' : увеличение концентрации добавок повышает среднюю прочность (критерий Джонкхиера): p-value $= 2.936 \times 10^5$.

Виды эффектов модели

Прежде чем переходить к вопросу, средние (медианы) в каких группах отличаются (в независимости от того, отклонили мы гипотезу однородности или нет), следует понять, чем вызваны различия между выборками. Наиболее популярными являются 2 модели: *модель со случайным эффектом*

$$X_{ij} = a_j + \varepsilon_{ij},$$

где $\{a_j\}$ – н.о.р. случайные величины (как правило, нормальные) со средним μ и дисперсией σ_α^2, независимые с $\{\varepsilon_{ij}\}$, и *модель с фиксированным эффектом*

$$X_{ij} = \mu + a_j + \varepsilon_{ij},$$

где a_j – не случайны.

Различия эффектов

ANOVA: переменные

Факторы

FIXED

Исследователь: Сравню-ка я эффективность анальгина и лекарства СтопБобо, под контролем плацебо!

Как бы ни было поставлено это исследование, группы будут три, и именно эти, **других нет**.

RANDOM

Исследователь: Изучу-ка я, различается ли масса лягушек в разных прудах!

Количество прудов в исследовании может быть разным, **существуют неисследованные** пруды.

Модель с фиксированным эффектом

Проверять гипотезы об однородности пар выборок внутри совокупности в модели со случайным эффектом бессмысленно, потому что различия будут вызваны случаем. Однако в модели с фиксированным эффектом такая задача интересна.

Свойства модели:
1) Разбиение на группы определено до получения данных.
2) При повторе эксперимента ожидается, что соотношения между средними групп сохранятся.
3) Если между средними есть различия, на следующем этапе анализируется, какие именно группы различаются.

Пусть $\{X_{ij}\}_{i=1}^{n_j} \sim N(\mu_j, \sigma^2), 1 \leq j \leq k$. Критерий проверяет гипотезы $H_{0j}: \alpha_j = \alpha_{j+1}$, где α_j снова упорядочены по возрастанию выборочных средних $\overline{X}_j$. Рассмотрим

$$LSD_j = t_{1-\frac{\alpha}{2}} \sqrt{\frac{n_j + n_{j+1}}{n_j n_{j+1}}} \sqrt{\frac{(n_j - 1)S_j^2 + (n_{j+1} - 1)S_{j+1}^2}{n_j + n_{j+1} - 2}}.$$

где t_γ – γ-квантиль распределения Стьюдента с $n_j + n_{j+1} - 2$ степенями свободы, S_j^2 и S_{j+1}^2 – выборочные дисперсии j-той и $(j+1)$-ой выборки соответственно.

Если $|\overline{X}_j - \overline{X}_{j+1}| > LSD_j$, то частная нулевая гипотеза $H_{0j}: \alpha_j = \alpha_{j+1}$ отклоняется в пользу двусторонней альтернативы. LSD можно использовать только в случае отклонения общей гипотезы однородности, и при этом стоит применять множественную проверку гипотез.

Критерий Неменьи

Непараметрический аналог критерия HSD Тьюки. Пусть в каждой из k выборок n наблюдений. Пусть R_{ij} – ранг наблюдения X_{ij} в общей совокупности, $\overline{R}_j = \frac{1}{n}\sum_i R_{ij}$ – средний ранг по j-той выборке.

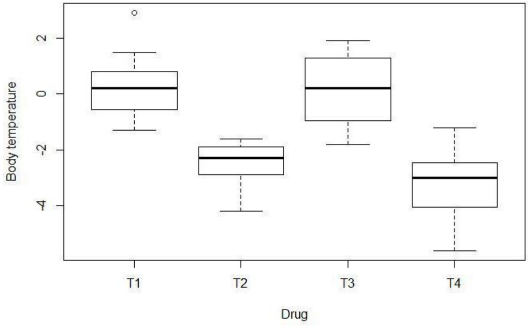
Введем

$$CD = q'_{1-\frac{\alpha}{2}} \frac{k+1}{6n},$$

где q'_γ – γ-квантиль из распределения стьюдентизированного размаха с k степенями свободы.

Проверим серию гипотез $H_{0j} : \alpha_j = \alpha_{j+1}$, где α_j упорядочены по возрастанию $\overline{R}_j$. Если $|\overline{R}_j - \overline{R}_{j+1}| > CD$, то отвергаем гипотезу H_{0j}. Для проверки H_{0j} следует пользоваться методами множественной проверки гипотез.

Действие жаропонижающих на морских свинок:



LSD Фишера

T_1 vs. T_3	0.9983
T_3 vs. T_2	3.5×10^{-8}
T_2 vs. T_4	0.2949

Критерий Неменьи

T_1 vs. T_3	0.9999
T_3 vs. T_2	1.8×10^{-4}
T_2 vs. T_4	0.7942

Двухфакторный дисперсионный анализ

Пусть имеются наблюдения признака X на N объектах. Хотим проверить, зависят ли значения признака X (а точнее, его среднее или медиана) от факторов A и B, принимающих значения $(A_1, \ldots, A_k)$ и $(B_1, \ldots, B_m)$ соответственно.

Пусть при $A = A_j$ и $B = B_l$ значения признака X заданы выборкой $\{X_{ijl}\}_{i=1}^{n_{jl}}, 1 \le j \le k, 1 \le l \le m$.

Поскольку двухфакторный анализ для выборок разного размера довольно сложен, будет считать, что $n_{11} = \ldots = n_{km} = n$. Часто будем полагать, что $n = 1$.

Двухфакторный дисперсионный анализ

Линейная двухфакторная модель:

$$X_{ijl} = \mu + \alpha_j + \beta_l + \gamma_{jl} + \varepsilon_{ijl},$$

$i = 1, \ldots, n$; $j = 1, \ldots, k$; $l = 1, \ldots, m$.

μ – глобальное среднее признака X;
α_j – воздействие j-того уровня фактора A;
β_l – воздействие l-того уровня фактора B;
γ_{jl} – дополнительное воздействие комбинации уровней j и l факторов A и B соответственно;
ε_{ijl} – н.о.р. случайные ошибки.

Двухфакторный дисперсионный анализ

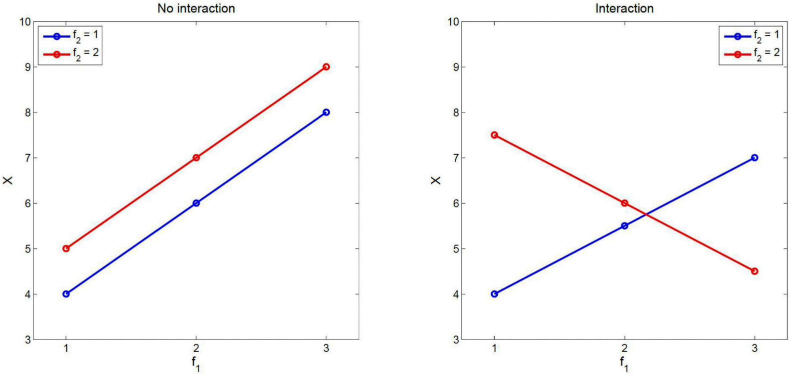
Если $\gamma_{jl} = 0 \ \forall j, l$, то решить задачу дисперсионного анализа гораздо проще (можно свести задачу к однофакторному дисперсионному анализу для связанных выборок). Иначе приходится рассматривать следующие гипотезы:

$H_0^1 : \alpha_j = 0 \ \forall j$ (т.е. фактор A не влияет на значения признака X) против $H_1^1 : H_0^1$ *неверна*,

$H_0^2 : \beta_l = 0 \ \forall l$ (т.е. фактор B не влияет на значения признака X) против $H_1^2 : H_0^2$ *неверна*,

$H_0^{12} : \gamma_{jl} = 0 \ \forall j, l$ (т.е. между факторами A и B нет взаимодействия) против $H_1^{12} : H_0^{12}$ *неверна*.

Пример: X – успешность решения кейса командой (в баллах от 0 до 10), f_1 – размер команды (1 – маленькая, 2 – средняя, 3 – большая), f_2 – наличие назначенного лидера (1 – нет, 2 – есть).



Пусть $X_{ijl} \sim N(\mu_{jl}, \sigma^2)$, $\mu_{jl} = \mu + \alpha_j + \beta_l + \gamma_{jl}$. Обозначим

$\overline{X}_{jl}$ – выборочное среднее по ячейке;

$\overline{X}_{j*}$ – выборочное среднее по значению фактора $A = A_j$;

$\overline{X}_{*l}$ – выборочное среднее по значению фактора $B = B_j$;

$\overline{X}$ – выборочное среднее по всей таблице.

Внутрифакторные дисперсии:

$$S_1^2 = \frac{nm}{(k-1)} \sum_{j=1}^{k} (\overline{X}_{j*} - \overline{X})^2, \quad S_2^2 = \frac{nk}{(m-1)} \sum_{l=1}^{m} (\overline{X}_{*l} - \overline{X})^2,$$

$$S_{12}^2 = \frac{n}{(k-1)(m-1)} \sum_{j,l} (\overline{X}_{jl} - \overline{X}_{j*} - \overline{X}_{*l} + \overline{X})^2,$$

$$S_{int}^2 = \frac{1}{km(n-1)} \sum_{i=1}^{n} \sum_{j,l} (X_{ijl} - \overline{X}_{jl})^2.$$

Нормальный двухфакторный дисперсионный анализ

Проверка значимости факторов и взаимодействия между ними:

1) при $n > 1$

$$F_1 = \frac{S_1^2}{S_{int}^2} \sim F(k-1, km(n-1)) \text{ при верной } H_0^1;$$

$$F_2 = \frac{S_2^2}{S_{int}^2} \sim F(m-1, km(n-1)) \text{ при верной } H_0^2;$$

$$F_{12} = \frac{S_{12}^2}{S_{int}^2} \sim F((k-1)(m-1), km(n-1)) \text{ при верной } H_0^{12};$$

2) при $n = 1$ (предполагаем, что гипотеза H_0^{12} верна)

$$F_1 = \frac{S_1^2}{S_{12}^2} \sim F(k-1, (k-1)(m-1)) \text{ при верной } H_0^1;$$

$$F_2 = \frac{S_2^2}{S_{12}^2} \sim F(m-1, (k-1)(m-1)) \text{ при верной } H_0^2.$$

Пример

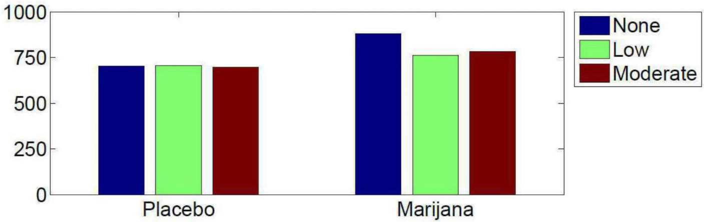
Изучалось воздействие марихуаны на скорость реакции. В качестве испытуемых были выбраны по 12 человек из каждой категории:

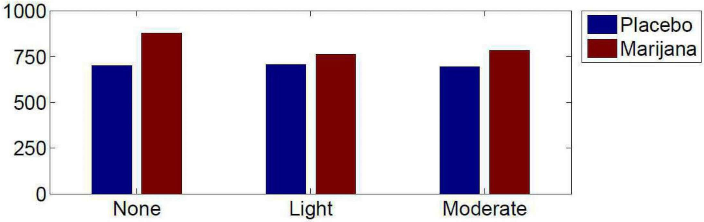
- никогда не пробовали марихуану;
- иногда употребляют марихуану;
- регулярно употребляют марихуану.

Испытуемые были разделены на две равные группы; половине из них дали выкурить две сигареты с марихуаной, вторая половина выкурила две обычные сигареты с запахом и вкусом марихуаны. Сразу после этого все испытуемые прошли тест на скорость реакции.

Требуется оценить влияние марихуаны на скорость реакции, учитывая фактор предыдущего опыта употребления.

Пример


1000
750
500
250
0
Placebo
Marijana
None
Low
Moderate


1000
750
500
250
0
None
Light
Moderate
Placebo
Marijana

H_0^1 : средняя скорость реакции одинакова при употреблении и марихуаны, и сигарет;

H_0^2 : средняя скорость реации не зависит от предыдущего опыта употребления марихуаны;

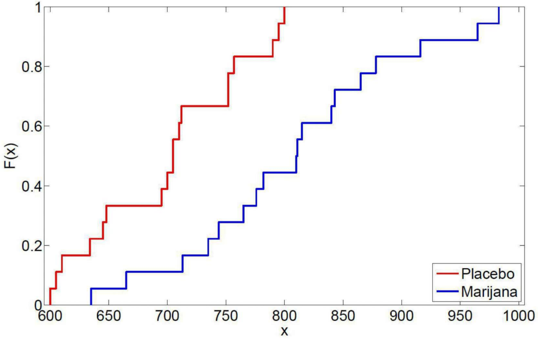
H_0^{12} : отсутствует межфакторное взаимодействие между употребляемым веществом и предыдущим опытом употребления марихуаны.

Source	F	p-value
Group	17.58	0.0002
Past use	2.02	0.15
Interaction	2.02	0.15

Пример

Вывод: гипотеза о том, что предыдущий опыт употребления не влияет на скорость реакции, не отклоняется – значит, данные по группам можно объединить.

Для объединенных данных:
1) p-value однофакторного дисперсионного анализа – 0.00036;
2) p-value критерия Манна-Уитни – 0.00059;
3) p-value двухвыборочного t-критерия – 0.00018.



Спасибо за внимание!